SYNTHETIC DATA GENERATION FOR THE OPTIMIZATION OF STRAINS IN METABOLIC ENGINEERING USING LATENT SPACE REPRESENTATIONS DERIVED FROM A CONDITIONAL VARIATIONAL AUTOENCODER (CVAE)

TU Delft

1. Background

- One crucial factor associated with engineering a strain is the **expense of generating data** used to steer the engineering process on an industrial scale
- The costs are related to a combinatorial explosion [1] in the screening process, resources to run experiments, and the amount of (Full-time equivalent) FTE. These costs are estimated to be upward of **€10,000.00 for an experiment**.
- To lower costs for such experiments we intend to explore **generative models to produce synthetic data** for metabolic engineering experimentation strategies by modeling the underlying distribution and structure of a dataset.
- **Variational Autoencoders (VAE)** are generative models that can be trained using conditional information to generate a synthetic dataset[2].

2. Research Question

Can a Conditional Variational Autoencoder be used to learn the underlying distribution of the data such that a reduced-dimensional latent space representation of a strain can be used to generate a synthetic dataset for combinatorial pathway optimization experiments?

Sub questions:

- How many **latent dimensions** do we need to capture the **distribution** and the **structure** of the dataset?
- How well does the **probabilistic PCA** capture the distributions and structure of the dataset as a baseline generative model?
- How well do the **VAE** and **CVAE** capture the distributions and structure of the dataset?

7. References

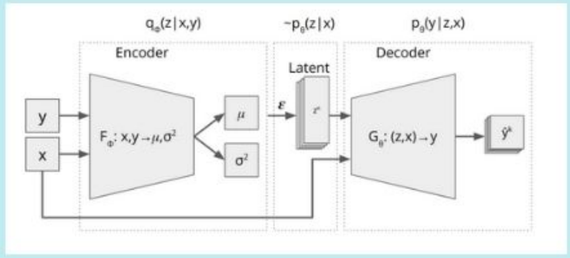
[1] M. Jeschek, D. Gerngross, and S. Panke, "Combinatorial pathway optimization for streamlined metabolic engineering," Current Opinion in Biotechnology, vol. 47, pp. 142–151, Oct. 2017, doi: 10.1016/j.copbio.2017.06.014.

[2] C. Doersch, "Tutorial on Variational autoencoders," arXiv.org, Jun. 19, 2016. https://arxiv.org/abs/1606.05908

[3] M. Debbagh, "Learning Structured Output Representations from Attributes using Deep Conditional Generative Models," arXiv (Cornell University), Apr. 2023, doi: 10.48550/arxiv.2305.00980.

3. Methodology

- The original dataset consists of **5000 simulated strains** generated from a kinetic model resembling E. coli. strains containing each **19 kinetic parameter** values.
- The quantitative metric was the **KL-div**[3] between the original and synthetic datasets. For visual inspection, the **PCA reduction** was plotted. A synthetic dataset, containing 5000 strains, is generated by the PPCA, VAE, and CVAE.
- A **comparative analysis** is carried out on the quantitative metric and visualization.

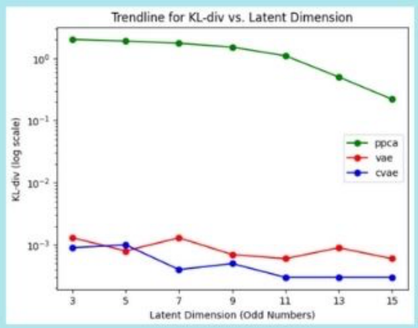


PPCA	n_components: 3, 5, 7, 9, **11**, 13, 15
VAE & CVAE	latent_dim: 3, 5, 7, 9, **11**, 13, 15 epoch: 100, 200, 300, **400**, mini-batch: 16, 32, **64**, 128 lr: 0.01, 0.001, **0.0001**, 0.00001 weight_decay: 0.1, 0.01, **0.001**, 0.0001

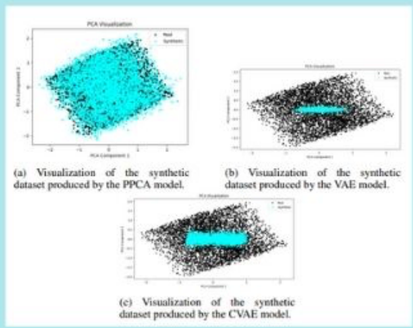
- For the PPCA we **vary the principal components** to establish how many we need to model the distribution of the dataset.
- For the VAE and the CVAE models we vary the **size of the latent layers** is varied to establish how many features we need to model the distribution of the dataset.

4. Results & Discussion

- For the PPCA model, we observe a downward trend in the KL-div values as the number of principal components used increases.
- For the VAE model, we observe very low values of the KL-div. The distribution of the dataset is already captured for smaller latent layer sizes.
- For the CVAE model, we observe the same low values for the KL-div. Conditioning the VAE shows improvement on the VAE model. However, noticeable differences are observed when measured with a precision of four decimal places



- The PPCA model doesn't place precise boundaries, but the general structure of the dataset is captured as observed using 11 principal components.
- The synthetic dataset from the VAE is observed to have produced values along the mean with minimal variance along the first component. This remains consistent for the latent layer size of 11 and higher.
- The CVAE produced values along the mean with minimal variance along the first component and second component. No difference is observed for the latent layer size of 11 and larger.



- When comparing the MSE measurements between the models, we observe that the VAE and the CVAE are prioritizing the modeling of the underlying distribution during training.
- The KL-div parts of the ELBO function as a larger influence than the MSE when it comes to training the models.
- We can add an extra parameter, namely the beta parameter, to the ELBO function to offset the influence of the KL-div during training.

$$\mathrm{KL}[q(z|X)\,||\,p(z)] = -\frac{1}{2}\sum_i \left(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2\right)$$

$$\mathrm{ELBO} = \mathrm{MSE}(X, \hat{X}) - \mathrm{KL}\left[Q(z \mid X) \,\|\, P(z)\right]$$

Latent	3	5	7	9	11	13	15
PPCA							
MSE	0.2177	0.1754	0.1366	0.0974	0.0598	0.0291	0.0131
KL-div	**2.0153**	**1.8956**	**1.7468**	**1.5181**	**1.0991**	**0.5015**	**0.2201**
VAE							
MSE	0.0586	0.0591	0.0586	0.0590	0.0594	0.0589	0.0593
KL-div	**0.0013**	**0.0008**	**0.0013**	**0.0007**	**0.0006**	**0.0009**	**0.0006**
CVAE							
MSE	0.0545	0.0541	0.0548	0.0548	0.0549	0.0550	0.0548
KL-div	**0.0009**	**0.0010**	**0.0004**	**0.0005**	**0.0003**	**0.0003**	**0.0003**

5. Conclusion

- The study compares PPCA, VAE, and CVAE for pathway optimization. PPCA, a statistical model, captures the underlying structure but requires a large latent space.
- The study compares PPCA, VAE, and CVAE for pathway optimization. PPCA, a statistical model, captures the underlying structure but requires a large latent space.

6. Future work & Limitation

Further collaborative efforts, as demonstrated in the CSE3000 Research Project, could yield valuable insights for advancing combinatorial pathway optimization in metabolic engineering.

Author: Neil Alwani, N.M.Alwani@student.tudelft.nl

Supervisor: Paul van Lent, Responsible Professor: Prof. Dr. Thomas Abeel